DOCUMENT RESUME

ED 301 928 EA 020 365

AUTHOR Brandt, Richard M.
TITLE Improving the Quality of Student Outcome Data Used in Career Ladder Promotion Decisions.
PUB DATE Apr 87
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Career Ladders; Elementary Secondary Education; *Evaluation Criteria; *Incentives; Merit Pay; *Outcomes of Education; Politics of Education; *Scores; *Teacher Evaluation; Teacher Salaries

ABSTRACT

Incentive pay plans for teachers are either planned or underway in 37 states; at least 10 states tie improvements in student performance to teacher evaluation criteria. Legislatures and school boards have mandated these criteria despite educational researchers' and test specialists' warnings concerning teachers' incomplete control over learning, lack of appropriate and valid measures, a wide variety of teaching assignments preventing common testing instruments, and varying student abilities and backgrounds. Since politicians and lay leaders remain unconvinced by these arguments, educators have taken at least two approaches: (1) standardized tests of basic skills; and (2) specific learning objectives aligned with selected measures. The Danville, Virginia, program uses student outcome data relevant to specific teaching objectives and other sources to judge career ladder candidates' teaching effectiveness. A previous paper discussed the nature of student outcome measures used during the study's first year. The present paper describes and assesses (1) the effectiveness of improvement efforts concerning instructional objectives cited in career ladder promotion plans; (2) measures used to gather student outcome data; and (3) data collection, scoring, and interpretation procedures followed in the Danville program. Out of 162 district learning goals, the vast majority (142) are cognitive in nature, and only 20 goals are either affective or behavioral. Results show local cumulative and published cumulative tests as more appropriate for assessing major year-long objectives. Results for administrating and scoring measures are complex and discussed at length. Included are several figures and six references. (MLH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED301928

Improving the Quality of Student Outcome Data Used in Career Ladder Promotion Decisions

Richard M. Brandt
University of Virginia

Presentation at AERA Convention, Washington, 4/87

INTRODUCTION

Incentive pay plans for teachers are in operation or under development in some 37 states. In at least 10 of them improvements in student performance are among the criteria to be used to indicate good teaching.[1] Legislatures and school boards have mandated their inclusion despite considerable warning from educational researchers and test specialists that such practices may be inappropriate.

Among the many reasons cited for not using student test results for judging the effectiveness of teaching are (a) the lack of control teachers have over more than a limited portion of all factors influencing learning, (b) the unavailability of appropriate, valid, and secure measures for many important educational objectives (c) a wide disparity in teaching assignments which prevents the use of common instruments across an entire school staff, and (d) varying student abilities and backgrounds from one class to another which make direct comparisons tenuous at best. Berk cites 19 sources of error that can be expected in the use of student test data for such purposes.[2] Loyd[3] and Millman[4] are others who list the pitfalls of this practice.

Yet politicians and lay leaders see student learning as the primary reason why schools exist and often cite anecdotes about teachers who have had profound, lasting impacts on their students. They are not convinced when they hear arguments that student measures are inappropriate for judging teachers.

Responding to these pressures, educators have taken at least two approaches to the use of student outcome measures. In one approach, standardized tests of basic skills have been used as one, and in some instances the only, basis for merit pay.[5] Regression models are typically established in which student gain scores are compared with expected gain scores predicted from pre-test or ability measures. Teachers of students who exceed expectancies are presumed to be superior. Problems of curricular validity, test security, and the stability of residual gain scores are particularly pronounced with regression models. At least half the teachers, furthermore, are only tangentially responsible for basic skill instruction, so there is an uneven application of the criteria at best.

EA 020 365

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard M Brandt
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

In the other approach, specific learning objectives are prescribed by or for teachers along with measures selected or designed to assess the amount of learning that hopefully will take place over a specified time period. While this goal-setting approach permits greater focus on each teacher's primary objectives, because grade and subject taught are appropriately considered, it reduces the use of common measures. Curricular validity is enhanced at the expense of (a) decreasing the use of common measures and (b) making comparisons more difficult. Many different instruments will be needed and the problems of cost, test security, test administration and interpretation are compounded considerably.

Despite the difficulties, however, certain advantages would seem to prevail in this approach over the regression model. In merit pay and career ladder programs it is most important that what teachers are expected to teach serves as the major focus for assessing how well they do it. Teachers should feel that what students are supposed to learn from them is what is going to be assessed, and not something else; otherwise many will feel they are at a disadvantage and the system is unfair. When teachers are allowed to specify the instructional goals they want to be assessed on, they have one less complaint about career ladder/merit pay plans. In fact, in a poll taken in Danville, Virginia, in April 1985, when a career ladder program was just being started, teachers chose student outcome data as the second most relevant source of data for determining teaching effectiveness out of several alternatives. Only the principal had a higher rating as a good source of data.

In the Danville program, now in its second year of implementation, student outcome data relevant to specific teaching objectives are being used along with other data sources (classroom observations, parent and student surveys, etc.) to pass judgment on the teaching effectiveness of career ladder candidates. Candidates not only specify the instructional goals on which pre- and post-student outcome data are to be collected but the means of measurement as well. Two separate committees review proposed objectives and measurement plans and after revision, if necessary, ultimately approve them as an acceptable component of the candidate review process. Findings are weighed at the end of the year in making final judgments about the effectiveness of teaching.

PURPOSE

The nature of student outcome measures used during the first year of the Danville program was presented in a previous paper.[6] The purpose of the present paper is to describe and assess the effectiveness of an effort to improve the nature of instructional objectives cited in career ladder promotion plans; measures used to gather student outcome data; and data collection, scoring, and interpretation procedures followed during the first and second years of implementation of a career ladder program. A summary of

psychometric deficiencies found in first year practices was used to provide improved guidelines and inservice training for those involved in the second year.

PROCEDURE

A coding system was developed to describe the specific nature of (a) instructional objectives included in the student outcome component (cognitive vs. affective/behavioral, single vs. multiple skill or content unit, short-term vs. long-term, etc.); (b) measures used (local unit or cumulative tests, published unit or cumulative tests, standardized tests, performance tasks, performance records, attitude inventories, behavior checklists, etc.); and (c) the precautionary steps taken in administering, scoring, and interpreting data to enhance valid collection and analysis of data.

A review of student outcome data used during the first year of the program revealed (a) a number of deficiencies and inconsistencies in the administering and scoring of measures and (b) relatively little basis for judging the significance of findings. In many instances, however, better measurement practices could have been followed by standardizing the instructions and response time for pre- and post-performance tasks; using disinterested second parties for scoring, administering or monitoring the administration of measures; selecting items for pre- and post-testing randomly from a common pool; not allowing teachers to see the exact items to be used; and comparing outcome information with relevant local or state norms. Many of the measures used by candidates were used elsewhere in the school system, but comparisons were seldom made to help with interpretation.

An analysis was made of first-year plans, using the coding system, and presented to the principals and other evaluators in August, 1986. Considerable discussion ensued about how various measures might be administered, scored, and interpreted in psychometrically sound ways. Although evaluation teams were not directed to code incoming plans or to use the forms and category definitions that had been developed, it became clear from later actions of candidate review committees, especially those of the Central Selection Committee which had final approval of plans, that improvements in student outcome measurement were expected. The majority of first drafts of second-year student outcome objectives and measurement plans were not approved. In many cases, questions were raised about the significance of an objective and whether standards were high enough. In other cases, procedural suggestions were made, such as having second parties administer, approve, or score measures.

Following final approval by the Central Selection Committee, student outcome plans of current career ladder candidates were coded and summarized in the same manner as plans of the previous year's candidates. A member of the Central Selection Committee

Committee also coded a third of the candidates' plans to provide an estimate of the reliability of the coding system. Although the percentage of agreement between the two sets of codings is 89.5, a phi coefficient of only 0.67 obtains. The percentage of total codings by both coders of measures and procedures specified in the plans that are identical is 73.2. While moderate agreement is suggested by these figures, increased specification of measurement practices to be followed may also be indicated, as well as increased awareness by the coders of the specific nature of particular measures.

RESULTS

Learning Goals

During two years of implementation 47 candidates have listed 162 distinct "learning goals," i.e., student outcome objectives for their students. The vast majority (142) are cognitive in nature, referring to academic knowledge or intellectual thought processes such as reading, writing, and calculating. They are often stated as (a) specific percentages of students who will achieve mastery of particular grade-level competencies designated in a curriculum guide or (b) average raw score or percentile gains on tests of one sort or another. The remaining 20 learning goals are affective or behavioral in nature, relating to feelings, attitudes, values, or behavior patterns. Examples include growth in (a) citizenship, sportsmanship, or other aspects or personal/social development, (b) the amount of recreational reading, (c) interest in particular subjects, and (d) participation in vocational club activities.

Measures

The types of measures used vary considerably from subject to subject and grade level to grade level. Ten categories were developed for classifying the primary measuring instrument being used in order to enhance recognition of its potential limitations, as well as its strengths, and consideration of its best use. Local unit tests, for example, as well as single skill or chapter tests, tend to focus on relatively brief amounts of skill or knowledge acquisition. They are appropriate for assessing specific competencies one at a time, often at the end of specific teaching units. As isolated measures they typically cover very little of a year-long curriculum, which often makes them inappropriate for the assessment of major curriculum goals. To be useful as measures of major goals they should be used as part of a whole battery of unit tests and at the end of several weeks of learning many other things so retention, not just acquisition, is assessed. Local cumulative and published cumulative tests are usually more appropriate for the assessment of major, year-long objectives.

Figure 1 shows the numbers of measures of each type that were used by each candidate group. The increase in local unit test usage in the current group was caused by a considerably greater amount of attention to how elementary students were doing on several school district competency batteries, especially in mathematics. It is not clear whether retention over a substantial period of time or mere acquisition was being assessed.

	1985-86	1986-87	Total
local unit test	7	17	24
local cumulative test	19	20	39
published skill or chapter test	1	4	5
published cumulative test	12	14	26
performance task	13	16	29
performance record	5	3	8
standardized test	7	9	16
attitude inventory	1	0	1
behavior checklist	8	1	9
other, including rating scales	1	4	5
Total	74	88	162

Figure 1. Types and frequencies of student outcome measures used by career ladder candidates

The relatively high use of performance tasks is explained by its appropriateness in the assessment of such learning as motor skills in physical education classes, typing skills in business education classes, and paragraph writing and composition in language arts classes. A performance task is defined as "a prescribed task in a standardized format so that students' performances under similar conditions can be observed, recorded, and compared validly . . . . The standardized format usually includes similar instructions, response time, and scoring procedures for all students."

Administration of measures

To enhance the collection of valid student outcome data in candidates' classes, four kinds of procedures were recommended for consideration and application as appropriate. The procedures and descriptions of each for coding student outcome plans, and their frequencies of usage follow:

	1985-86	1986-87
Teacher not know items A teacher does not know what specific questions are included in the measure of student outcomes. This condition prevails even if a teacher has seen a much larger pool of items from which the questions will be selected.	0	1
Administered by other Someone other than the students' teacher administers the measure--some neutral person.	17	29
Administration monitored The teacher administers the measure with the principal or some other neutral person present.	29	39
Standardized conditions These conditions prevail when student response time, instructions, and context conditions are specified and similar from one administration of a measure to others against which comparisons will be made, e.g., pre-post achievement testing of the same group or testing several classes of a given grade.	14	34

As is readily apparent, there were substantial increases in the use of all administrative procedures except for the first kind, in which only one example was found. Although several teachers used the Science Research Achievement tests or subtests, this battery has been used annually in the elementary grades in Danville. There is no assurance that teachers do not know at least some of the items, so this category was not automatically checked when an SRA test was specified.

Scoring of measures

To enhance valid scoring of student outcome data five kinds of procedures were recommended for consideration and application as appropriate. The procedures and descriptions of each for coding student outcome plans, and their frequencies of usage follow:

	1985-86	1986-87
Monitored objective scoring This condition prevails when the teacher scores student outcome data objectively either under direct observation of another adult or with a second person verifying scoring accuracy by random spot-scoring the data.	17	21
By second party Scoring of the measure is done by a neutral, second person, rather than the teacher.	50	54
Blind scoring This condition prevails when the person scoring data does not know whether they are pre- or post-instruction scores, candidate- or noncandidate-data. It also exists in the scoring of objective, standardized tests.	7	18
Multiple judges This condition refers to the use of two or more judges' scores in deriving a final overall score of student performance. It refers also to the use of two or more qualified persons establishing scoring standards or approving a teacher-made local test as appropriate. The need for multiple judgments is greater when scores take the form of high inference ratings or rankings.	5	18
Other Other conditions that enhance scoring accuracy and objectivity. The use of computers or outside test scoring centers is included here.	10	16

Major increases are apparent in the use of blind scoring practices and multiple judges. Blind scoring practices included using numbers instead of names on papers to hide the identity of students while their papers were being scored. For blind scoring, the dates of student work samples should also be hidden so the scorer could not tell whether pre- or post-test data were being coded. Most plans did not specify this latter practice, so coding these two sets of data was less rigorous than it might have been.

In addition to the five scoring categories, ratings were made of the objectivity of the scores derived from each measure on a three-point rating scale: high, moderate, and low

objectivity for low inference, moderate inference, and high inference judgment respectively. Results were as follows:

	1985-86	1986-87
High objectivity	44	50
Medium objectivity	22	20
Low objectivity	7	0

Interpretation of measures

To derive meaning from data findings, it is often helpful to compare data collected with similar data from other classes; other teachers; or school district, state or national norms. It is also important to consider the ability levels of students involved and other context factors that influence learning. The use of these interpretation procedures during these two years is presented below:

	1985-86	1986-87
Within school comparisons Final judgments about the significance of student outcome results are based, at least in part, on comparisons with similar data from other classes in the same school. Comparison classes may consist of those taught by this teacher previously or concurrently.	9	18
School division comparisons Final judgments about the significance of student outcome results are based (a) at least in part, on comparisons with classes in other schools in the school division, or (b) on school division standards on well-prescribed, objectively scored criterion tests.	20	21
State/national norms Final judgments about the significance of student outcome results are based, at least in part, on comparisons with state or national norms.	4	9
Pre- post-random items Tests or test items used to assess student gain are selected randomly from a common pool of tests or items for placement in pre- and post-measures. This procedure represents a refinement of one aspect of the standardized condition category.	4	5

	1985-86	1986-87
Relevant factors Interpretations take into account such potentially influencing factors as the ability levels and background experiences of students when comparing groups and/or estimating the extent of real learning.	4	14

Planning Improvement

The frequencies of coded administrative, scoring, and interpretation procedures specified in the plans of the first and second candidate groups are presented in Figure 2. Theoretical frequencies, in parentheses, were derived from the proportion of measures used in 1985-86 versus 1986-87.

Year	No. of Candidates	No. of Measures	Administration Procedures	Scoring Procedures	Interpretation Procedures	Total Procedures
1985-86	21	74	60 (74)	89 (99)	41 (49)	190 (222)
1986-87	26	88	103 (89)	127 (117)	67 (59)	297 (265)
Total		162	163	216	108	487
Chi square =			4.85	1.87	2.40	8.47
p			< .05	NS	NS	< .01

Figure 2. Frequencies of procedures included in the collection and analysis of student outcome data.

Significant increases are evident in both the administrative procedures alone and the total procedural steps undertaken to enhance the validity of student outcome data and interpretation. Increases occurred, also, in the number of recommended scoring and interpretation procedures to be followed by the second-year candidates but not to a significant extent. Whereas the number of measures was up 19 percent in almost direct proportion to the number of candidates, increases in recommended procedures were considerably greater in all three types of procedures as well as the total (Administrative 72%, Scoring 43%, Interpretation 63%, Total 56%).

DISCUSSIONS

Despite a significant improvement in steps taken to enhance the validity of student outcome data being collected and used, along with other data sources, to assess teaching effectiveness, there is still considerable room for improvement.

Several types of recommended procedures were seldom or almost never pursued. In all but one instance, for example, candidates either knew precisely or had a reasonably good knowledge of the specific items included on the measures. They are frequently included in the textbook materials and curriculum guides that teachers use. Teaching to the task is often a legitimate primary objective, in fact. What is of concern here, is the possibility of teaching only the specific items that will be on a test rather than the more general content, skill or behavior they are supposed to reflect. To the extent that teachers know the latter and not the specific items to be included on a measure, more valid assessment of real learning goals is likely. In the design or selection of local tests and performance tasks especially, the actual items to be included might be drawn by a third party or by random selection from a larger pool of possible items. Tests or test items used to assess student gain, in particular, might be selected randomly from a common pool of tests or items for placement in pre- and post-measures.

Although the standardized conditions category was coded quite frequently in 1986-87 (34 times), it was not always clear whether instructions, student response time, and other context factors were indeed similar from one administration to another. If published or standardized tests were used, coders assumed, and perhaps wrongly, that standardized conditions prevailed. The same was true of division competency tests which tend to be quite precise in how they are administered and scored. Nevertheless, important variations can occur in the administration of standardized tests, and it is important that conditions be well specified ahead of time and closely monitored if fair and comparable assessment practices are to be assumed. In a number of cases post-measures, but not pre-measures, were administered or monitored by a second party; both should be so administered or monitored to preclude the possibility of an artificially depressed baseline.

Another promising practice seldom observed so far is a completely blind scoring of data so that not only are students' names not apparent to the scorer but the date of performances. It is particularly important when scores are derived from subjective judgments, as contrasted with objective scoring keys, that both pre- and post-measure responses be scored blindly and at the same time.

Considerable improvement was noted in the use of records of previous classes taught by candidates or of similar classes in the same schools to help establish appropriate learning goals or to interpret the significance of outcomes. As stipulated in the Danville guidelines, results of student outcome data collection will be meaningful to the extent that comparisons can be made with relevant reference groups in the same or other schools or with local, state, or national norms. While there was also greater mention this year (N=14) of such relevant factors as the ability level of students for setting goals or interpreting results, there is still considerable room for improvement in the consideration of both student ability and other background factors.

Several of the learning goals targeted low ability groups, such as those who scored in the lowest quartile of the SRA reading test. Some of these goals were stated as an average percentile increase for a specified percentage of this group. No consideration was given to the regression effect that might automatically elevate some of those in this lowest quartile or of the need to track other students who might on subsequent testing take their place. A better goal statement might stipulate, as did one teacher, that the average of the percentage of items correct on the grade-level math competency test would be five percentage points higher than the average of the same students during the previous year. All students would be included in such a goal statement including those who might start out in the lowest quartile as well as those who might fall into it later.

Despite the need for still greater improvement in the kinds of measures used to assess student learning and procedures to be followed to enhance the validity of data, the major finding to be noted in this study is the significant improvement that has already taken place. Plans of second-year career ladder candidates are considerably better than those of their predecessors.

We assume, in the absence of a control group, that the development of a coding system for analyzing student outcome measurement plans and the attention given to valid measurement procedures in last summer's workshop and by the fall review committees, especially the Central Selection Committee, made the difference. Continued use of the coding system and attention to sound measurement practices will be needed if student outcome data are to provide useful, valid information for the promotion of career ladder candidates in the years ahead.

At least three specific activities can be suggested for improving goal-setting processes even more next year: (a) Workshops could be held at the beginning of the year for candidates as well as evaluators to study and code goal-setting plans of previous candidates. (b) Local norms could be developed for regularly used local, state and textbook tests in a variety of content areas by reviewing and organizing student test results already on file. They are needed both for establishing realistic

yet significant goals in the fall and interpreting data collected during the year. (c) Reviewing committees should formally code incoming plans before giving final approval to enhance their awareness of what is being proposed in goals and measures and what more might be done to strengthen the process. The use of student outcome data in the assessment of teaching is a defensible practice only to the extent that important, realistic, and well-defined learning goals are established and valid measurement practices followed.

References

1. "1986--Incentive programs for teachers and administrators: How are they doing?" Career Ladder Clearinghouse. Atlanta: Southern Regional Education Board, December 1986.

2. Ronald A. Berk, "The use of student achievement test scores as criteria for allocation of teacher merit pay," invited paper, March 22, 1984.

3. Brenda Loyd, "Measurement and statistical issues in the use of student outcome data," paper, AERA Convention, 1986.

4. Jason Millman, "Procedures for converting student achievement data into judgments about teacher effectiveness," paper, AERA Convention, 1986.

5. ERS Reports, Merit Pay for Teachers (1979); Merit Pay Plans for Teachers (1983). Arlington, VA: Educational Research Service.

6. Richard M. Brandt and Edward A. Kingston, "Individualized student outcome measures for career ladder candidates," paper, AERA Convention, 1986.